OVERVIEW OF RECOMMENDED MODIFICATIONS OF THE TOP-TO-BOTTOM METRIC TO IMPROVE IDENTIFICATION OF FOCUS SCHOOLS

Context

At the December 19, 2012 meeting of the Education Alliance at the Michigan Department of Education (MDE), it was determined that it was desirable to modify the top to bottom metrics to blunt the impact of outliers on the identification of focus schools. It was further determined that it was desirable to blunt the impact of positive outliers (very high scoring students) as well as negative outliers (very low scoring students).

There were both statistical and policy rationales for blunting the impact of outliers on both ends. The statistical rationale was that there is more measurement error (or noise) in both the positive and negative ends of student score distributions, and that blunting the impact on both sides is desirable to minimize the impact of poorly estimated achievement whether the poorly estimated achievement is on the top or bottom end.

The policy rationale was that focus identification may inappropriately influence school configuration decisions. For example, housing a gifted and talented program within a school may bring up the top 30 group scores sufficiently to identify such schools as focus schools. On the other end, housing an alternative education or special education center program within a school might bring the bottom 30 group scores down enough to identify such schools as focus schools. Blunting the impact of outliers on both ends would allow for school configuration decisions to be based on educational concerns rather than on concerns about impacts on accountability designations.

MDE’s Bureau of Assessment & Accountability (BAA) committed to proposing approaches to blunting the impact of outliers, and taking those proposed approaches to the BAA’s Technical Advisory Committee (TAC) and to BAA’s Advisory Committee (AC). The BAA TAC is a group of nationally recognized technical experts in psychometrics, statistics, and measurement. The BAA AC is an advisory group of stakeholders representing education associations, ISDs, and higher education that is more focused on policy issues. BAA further committed to receiving feedback and recommendations from the TAC and AC to take back to the State Superintendent, and ultimately to the Education Alliance association heads for their support.

Meeting with Technical Experts Chosen by the Education Alliance

Following the December 19, 2012 meeting, BAA staff met with the technical experts brought to the meeting by the Education Alliance association heads to discuss possible methods of blunting the impact of outliers on the identification of focus schools, at both the lower end and the upper end. At that meeting, two broad concepts were put forward. They were:

1. Normalizing the student z-score distributions to eliminate extreme outliers and to make the impact of positive and negative outliers symmetrical.
2. Capping the student z-score distributions to blunt the impact of large positive and large negative z-scores.

Several possibilities for capping the z-scores were discussed. It was determined that tying the z-score caps in some way to Michigan's cut scores was desirable. One suggestion was to tie the z-score caps to the advanced cut scores. The rationale for choosing the advanced was to ensure that there still remains an incentive to move students who have achieved proficiency to still higher levels of achievement.

Another suggestion was to tie the z-score caps to the proficient cut score. The rationale for choosing the proficient cut score was to reflect that achieving proficiency is the bar that schools are asked to help all students reach.

Two options were discussed regarding caps on the top end. It was suggested that the caps could either be the same for every grade, subject, and test combination or they could differ by grade/subject/test combination depending on the cut score or each combination.

Two options were also discussed regarding caps on the bottom end. It was suggested that the caps on the bottom end could be either the negative of the caps on the top end (e.g., the caps on the bottom and top end could be symmetric) or they could be set independently of the caps on the top end.

BAA Deliberations

After the meeting with the technical experts brought by the Education Alliance to the December 19, 2012 meeting, BAA staff deliberated on the pros and cons of each suggestion.

Normalizing the Student Z-Score Distributions

There were no identifiable cons to normalizing the student z-score distribution. Therefore, student scores were transformed into normalized z-scores using the following steps for each grade/subject/test combination.

1. Order unique observed scores in ascending order.

2. Obtain the frequency of each unique observed score.

3. Calculate the percentile rank of each unique observed score as:

$$PR_J = 100 * \frac{F_{<j} + F_j \quad 2}{N}$$

where

PR_J is the percentile rank of the j^{th} unique observed score,
$F_{<J}$ is the cumulative frequency of all unique observed scores with values less than the j^{th} unique observed score,
F_J is the frequency of the j^{th} unique observed score, and
N is the total number of observed scores.

This results in percentile ranks being in the (0, 100) range, non-inclusive, which allows for step 4 to function appropriately.

4. Calculate normalized z-score of each unique observed score as

$$z_j^* = \varphi^{-1}\ PR_j\ \ 100$$

where

z_j^* is normalized z-score of the j^{th} unique observed score, and
φ^{-1} is the inverse of the standard normal cumulative frequency distribution.

However, BAA’s large-scale data manipulation package (Microsoft SQL) does not have a function for φ^{-1}. To closely approximate φ^{-1}, BAA staff instead used a lookup table of percentile ranks running from 0.005 to 99.995 in increments of 0.01 with corresponding z_j^*s as excerpted in table 1 below.

Table 1. Lookup table translating percentile ranks into approximate normalized z-scores.

PR	*z**
0.005	-3.891
0.015	-3.615
0.025	-3.481
0.035	-3.390
…	…
49.975	-0.001
49.985	0.000
49.995	0.000
50.005	0.000
50.015	0.000
50.025	0.001
…	…
99.965	3.390
99.975	3.481
99.985	3.615
99.995	3.891

The z_j^*s were closely approximated by finding the percentile rank in the table nearest to PR_J and using the corresponding $z*$.

This procedure is able to flawlessly transform a radically non-normal continuous distribution into a normal distribution. For example, it was able to transform continuous log-normal distribution shown in the left panel of figure 1 below into the continuous normal distribution shown in the right panel of the same figure.

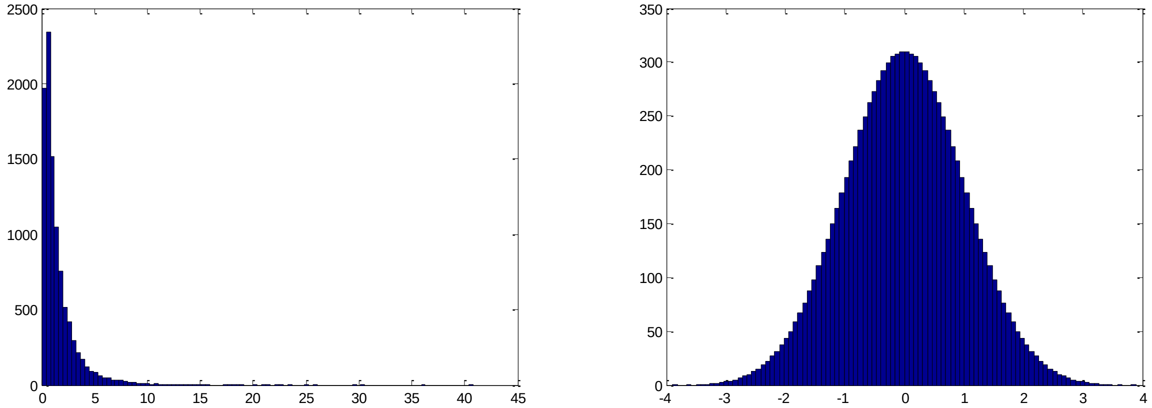


Figure 1. Lognormal distribution and normalized distribution.

For discrete distributions such as those resulting from state assessments, the procedure works well, but it not able to exactly normalize the distributions. Rather, it approximately normalizes the distributions. For example, in grade 3 MEAP mathematics and in MME mathematics, the non-normalized distributions of student scores are as shown in figure 2.

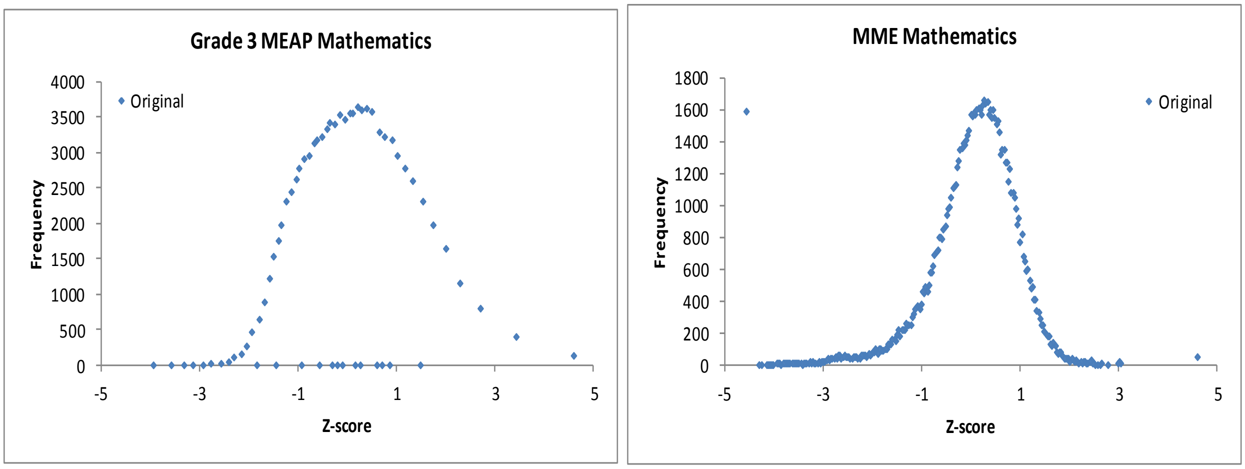


Figure 2. Non-normalized Grade-3 MEAP and MME mathematics distributions.

In figure 2, it is clear that the distributions are not normal. Rather, grade 3 MEAP mathematics is skewed to the right, and MME mathematics is skewed to the left, with a spike (nearly 1600) in students scoring the lowest possible score. When the normalizing procedure is applied to the data, it results in the distributions represented by the red dots in the figure 3. The resulting distributions are clearly more symmetrical than the original distributions. In addition, the cumulative frequency distributions of the normalized scores lines up nearly exactly with the cumulative frequency of the standard normal density, indicating that the normalizing transformation was successful.

One of the concerns raised by the TAC was that of the spike at the lower end on MME distributions, and whether that would still result in inordinate impacts of outliers on identifying focus schools. Because of the spike of nearly 1600 students achieving the lowest possible score, it is clear that normalizing alone is not sufficient to address the impact of outliers, and that capping is also needed. When capping is applied, there is nearly exactly the same number of students at the upper cap as at the lower cap, even with the spike seen on the MME graphs.

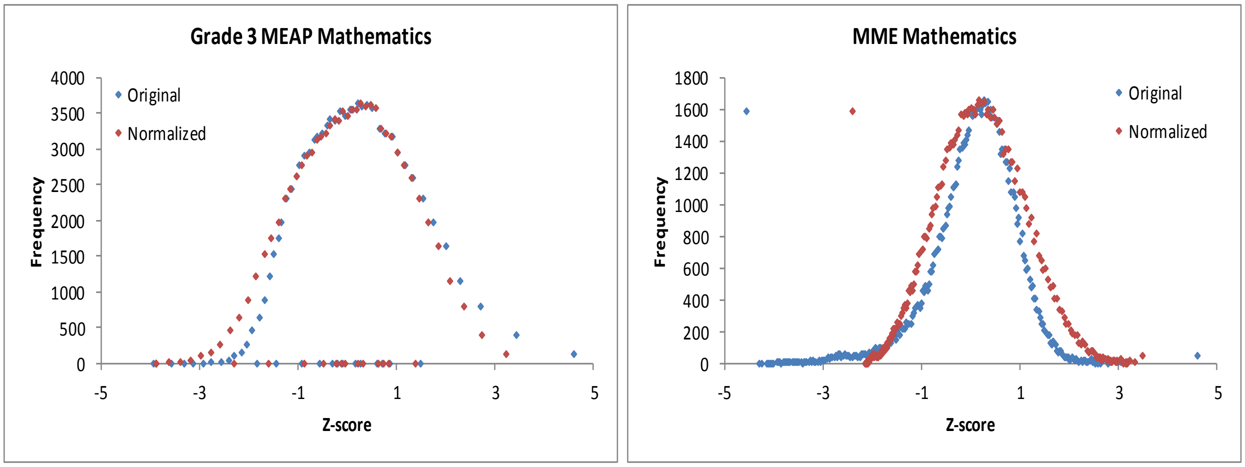


Figure 2. Normalized Grade-3 MEAP and MME mathematics distributions.

Capping the Student Z-Score Distributions

There were also no cons to capping the z-score distributions at some level. However, there were significant drawbacks to the different methods of identifying caps.

For caps on the upper end of student z-score distributions, the pros and cons of using different caps for each subject/grade/test combination follow. The pro of setting different caps for each combination would result the caps being tied directly to the cut scores for each specific subject/grade/test combination. The cons of such an approach are (1) that it would be difficult to explain that each combination is capped differently, and (2) that the subject areas with the highest cut scores would be less affected by the caps. Number (2) would result in the combinations with the highest caps driving the focus designation because greater variation would be allowable in those subjects. Because science and social studies have the highest cut scores, this would result in the focus designations being based largely on science and social studies, but only minimally on mathematics, reading, and writing. Because of unintended consequences this could produce, it was considered such a significant drawback that it was determined to take to the BAA TAC and BAA AC only those options in which the caps were set at the same level for each subject/grade/test combination.

There were also similar drawbacks to the different methods of identifying caps for the lower end of the student score distributions. BAA staff could think of no reasonable rationale for why the lower caps should not be symmetrical to the upper caps. For example, if the lower caps were allowed be further from the mean than the higher caps, then variation including greater measurement error on the lower end would largely drive focus designations. Conversely, if the upper caps were allowed be further from the mean than the lower caps, then variation including greater degrees of measurement error on the upper end would drive focus. BAA staff were unable to identify any reasonable rationale for allowing this to occur. Therefore, it was determined to take to the BAA TAC and BAA AC only those options in which the upper and lower caps were symmetrical.

To select possible cap locations, a simple set of analyses were run. After normalizing each z-score distribution, the normalized z-scores associated with the proficient and advanced cut scores were submitted to descriptive analysis. The results showed the following:

1. The maximum normalized z-score associated with an advanced cut score was 1.966.
2. The mean normalized z-score associated with an advanced cut score was 1.425.
3. The maximum normalized z-score associated with a proficient cut score was 1.015.
4. The mean normalized z-score associated with a proficient cut score was 0.173.

Because values from numbers 1, 2, and 3 (above) happened to be near the round numbers 2, 1.5, and 1, BAA staff reran the top to bottom ranking along with priority and focus designations in the following five ways to show the impact of each possible set of modifications:

1. Without any modifications.
2. Using normalized student z-scores without capping.
3. Using normalized student z-scores with caps at -2 and 2.
4. Using normalized student z-scores with caps at -1.5 and 1.5.
5. Using normalized student z-scores with caps at -1 and 1.

The results of these five runs were then taken to the BAA TAC meeting for review and recommendation.

BAA TAC Meeting and Recommendations

At the BAA TAC meeting, the TAC members were briefed on the issues behind the proposed modifications, and on the five options being investigated. The task for the BAA TAC was identified as providing recommendations to BAA on the proposed changes with the following guiding principles:

- Modifications should address the concerns about outliers having an inordinate impact on the identification of focus schools.

- Modifications should not result in a significant shift in the population of schools identified as priority schools (as the priority list is reasonably established and is not facing the type of criticism that is being leveled at the focus list).
- Modifications should not result in a total shift in the population of schools identified as focus schools (as the issues with the focus list is an inordinate impact of outliers on identification of schools as focus schools)
- Modifications should not result in a focus list that simply identifies the next lowest performing schools after priority schools (as the purpose of the focus metric is to identify the largest gaps rather than to identify low achieving schools).
- Modifications should not result in over identifying specific types of schools other than those that have large achievement gaps (e.g., should not result in focus school designation becoming a proxy for economic diversity).

The TAC was shown the scatterplots in figures 4-7 to demonstrate the impact of the modifications on top to bottom (TTB) rankings and on priority identification. In these scatterplots, the TTB percentile ranks for each option are compared to the original TTB percentile rank. Figure 4 shows that normalizing alone does not much affect TTB percentile ranks, as the correlation between the originals and those based on normalized data without caps is 0.9934. Figure 4 shows that normalizing and capping at -2 and 2 is similar, in that the correlation is 0.9930. Figure 5 shows that capping at -1.5 and 1.5 has more of an impact on TTB ranking and priority designation in that the correlation drops to 0.9884. Finally, figure 6 shows that capping at -1 and 1 has an even larger impact, with the correlation dropping to 0.9648.

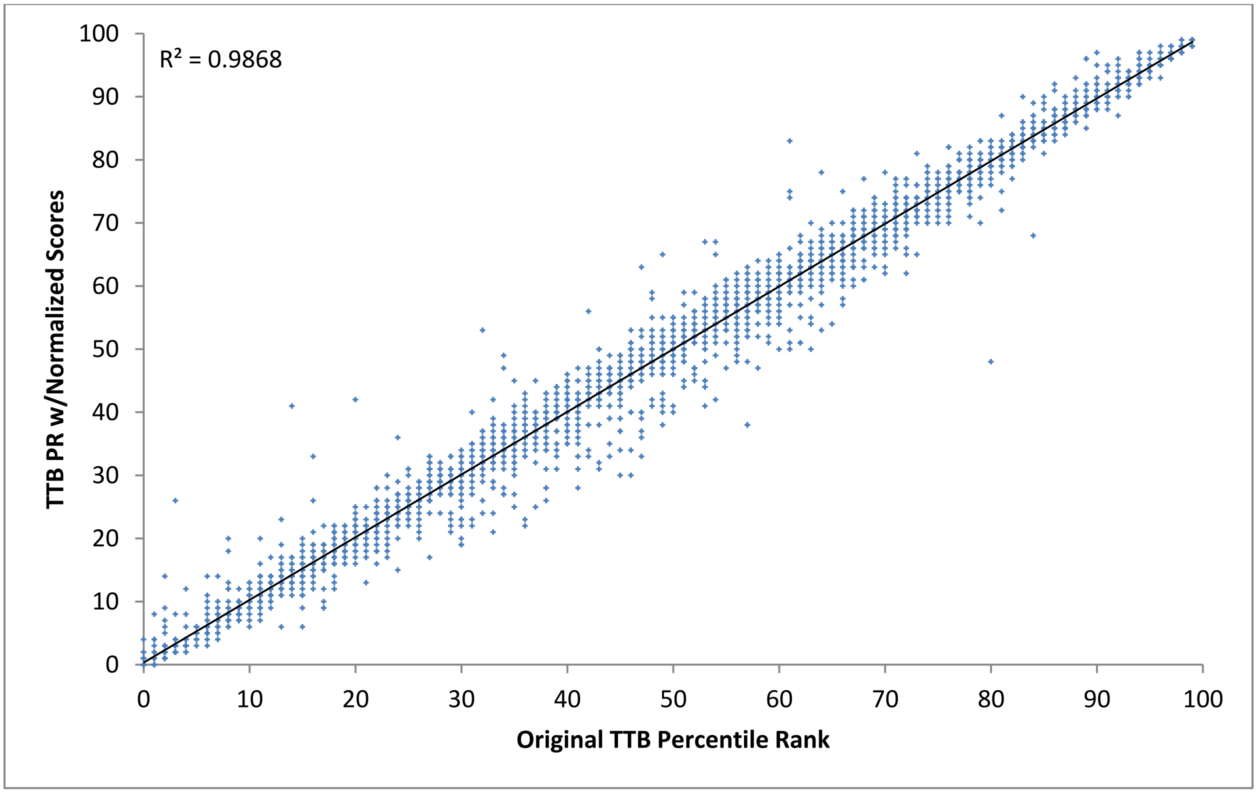


Figure 4. Relationship between original TTB ranks and TTB ranks based on normalized data without caps.

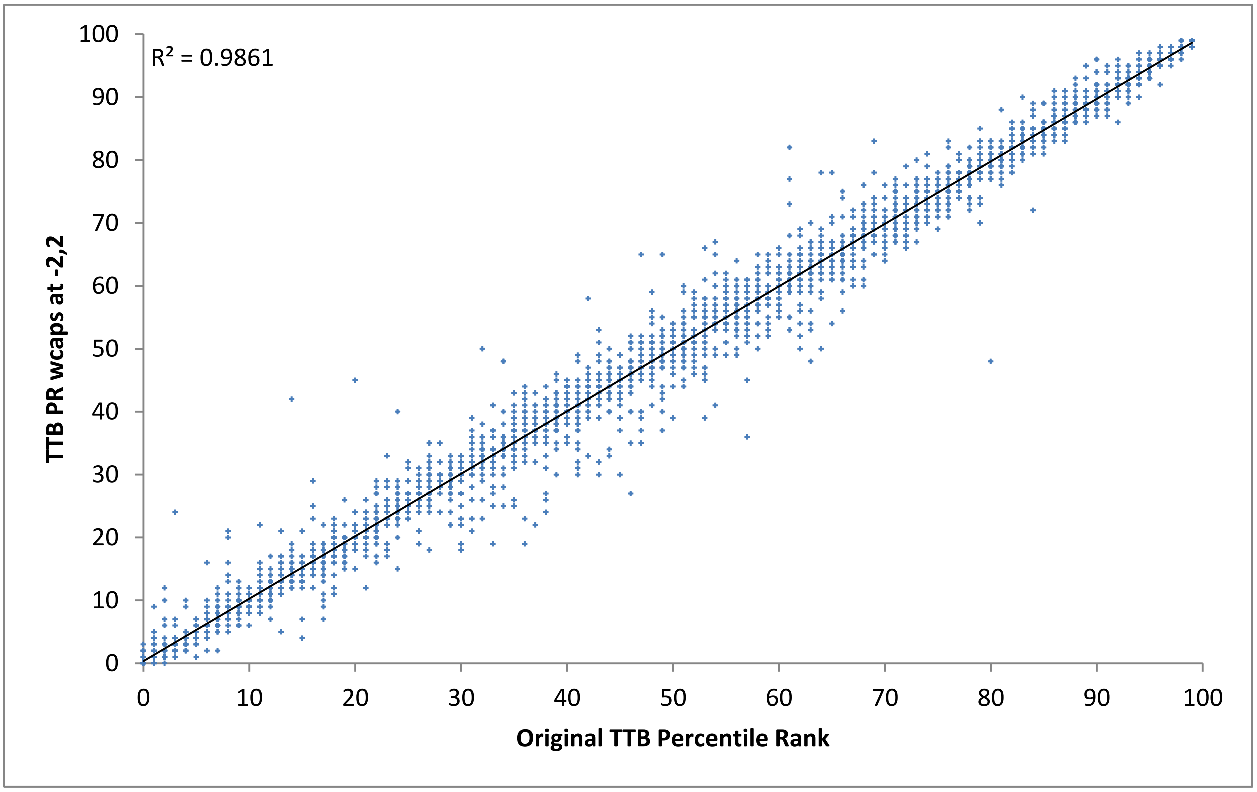


Figure 5. Relationship between original TTB ranks and TTB ranks based on normalized data with caps at -2 and 2.

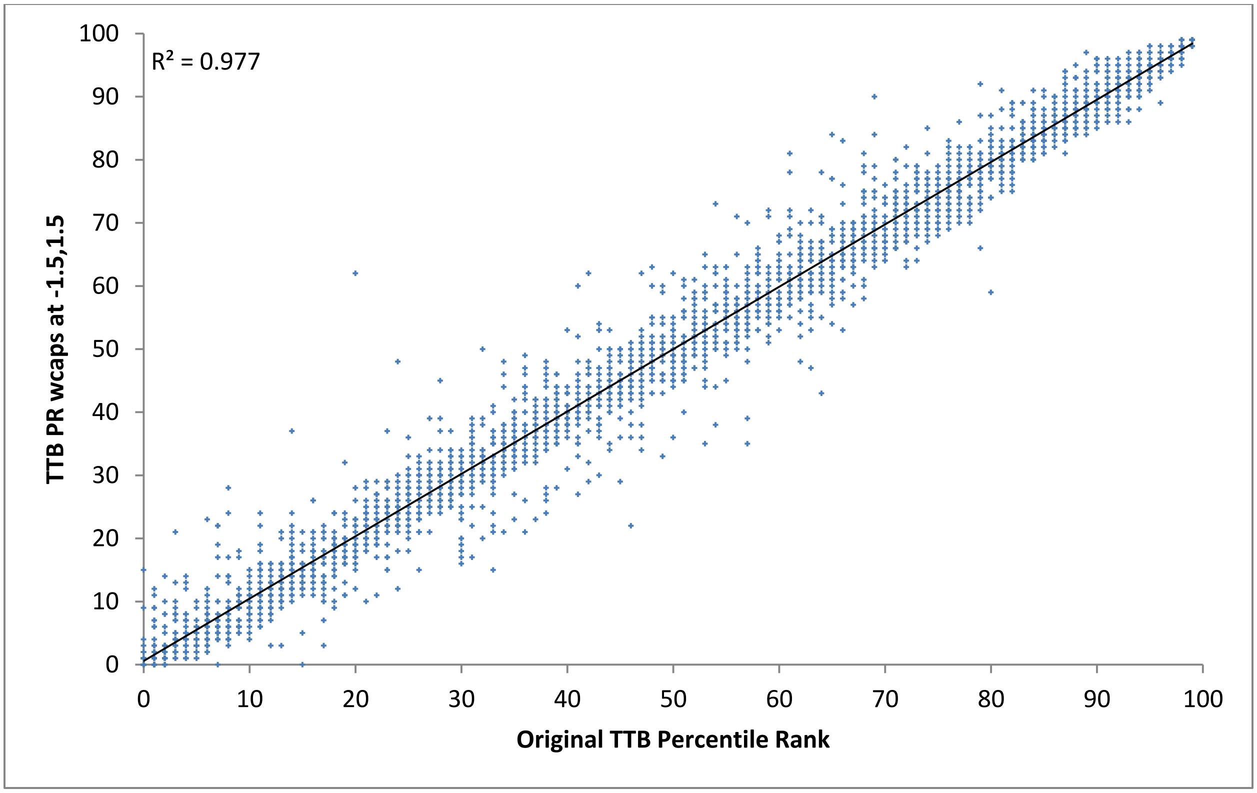


Figure 6. Relationship between original TTB ranks and TTB ranks based on normalized data with caps at -1.5 and 1.5.

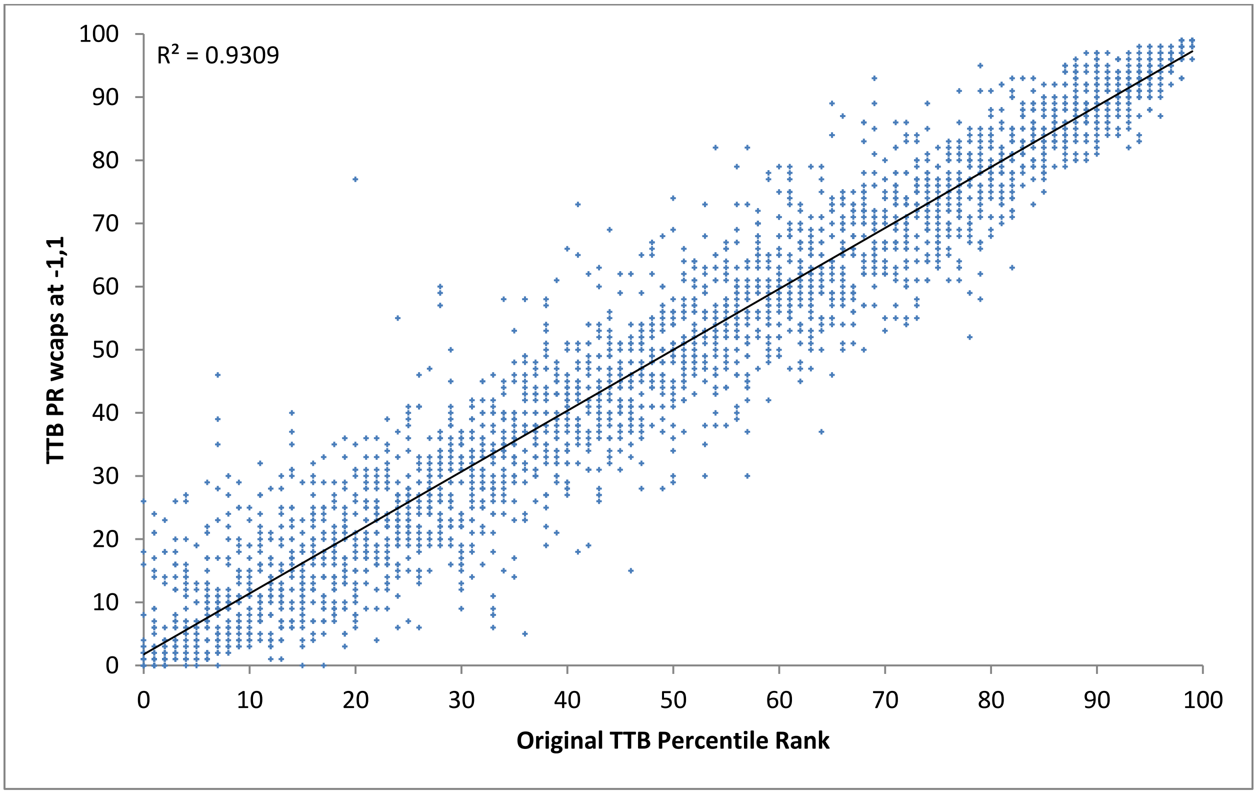


Figure 6. Relationship between original TTB ranks and TTB ranks based on normalized data with caps at -1 and 1.

In addition, the number of individual schools whose priority designation is affected by each option are presented in Table 2, with those whose focus designation is affected presented in Table 3.

Table 2. Consistency of priority designation with original.

	Impact on Priority Designation		
Modification	In original, Not in modified	In modified, not in original	In both
Normalized, no caps	10	9	136
Normalized, caps at -2, 2	16	15	130
Normalized, caps at -1.5, 1.5	42	42	104
Normalized, caps at -1, 1	57	58	88

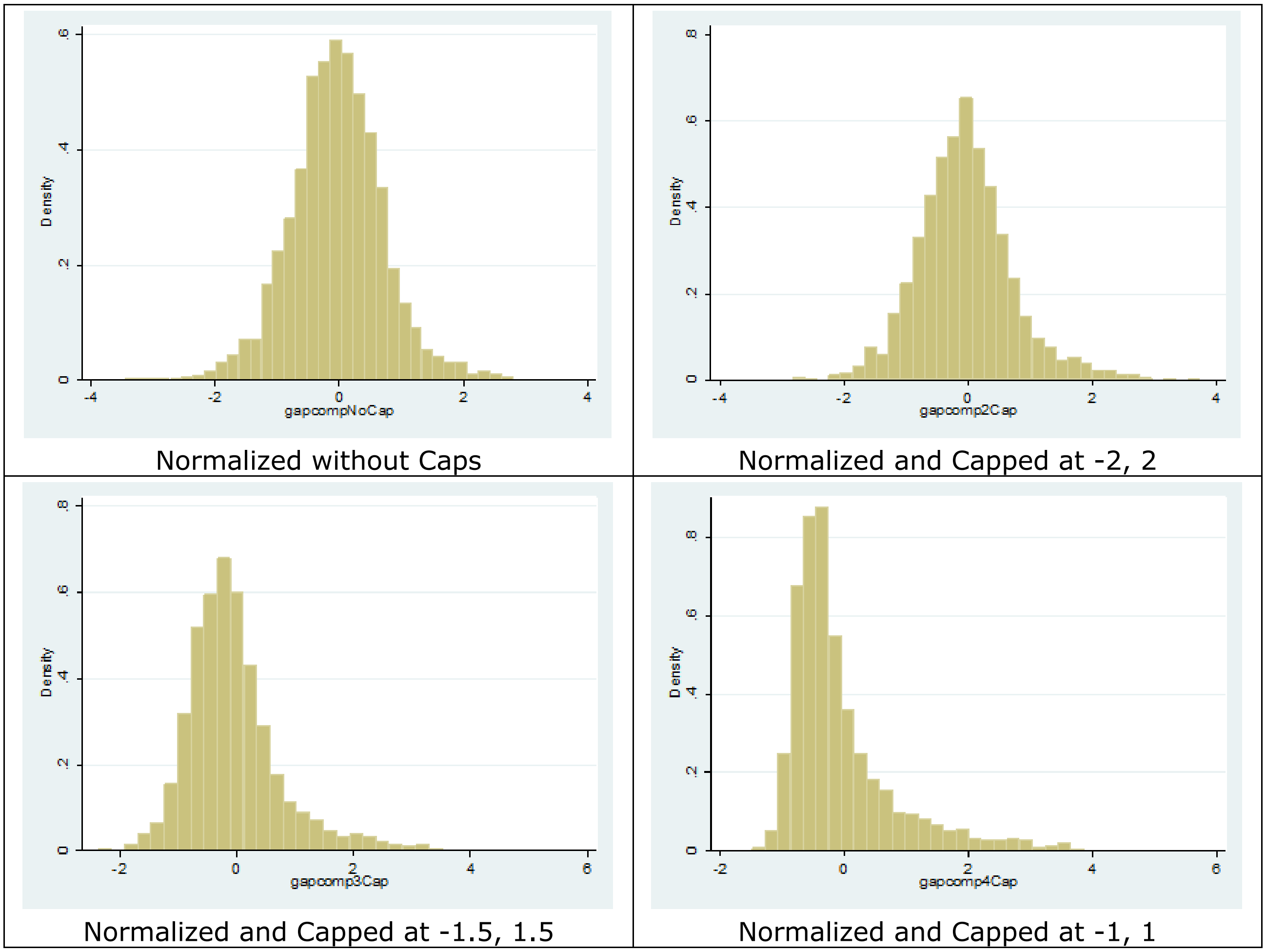
Table 3. Consistency of focus designation with original.

	Impact on Focus Designation		
Modification	In original, Not in modified	In modified, not in original	In both
Normalized, no caps	97	80	261
Normalized, caps at -2, 2	113	86	245
Normalized, caps at -1.5, 1.5	153	111	205
Normalized, caps at -1, 1	203	144	155

As can be seen in Table 2, priority designations do not shift much from the original with normalizing alone or with normalizing and placing caps at -2 and 2. However, with caps at -1.5 and 1.5, the impact results in nearly as many schools changing priority designation as those that are consistently classified as priority. Finally, capping at -1 and 1 results in more schools changing priority designation than those that are consistently classified as priority.

As can be seen from Table 3, the modifications have a greater impact on focus designation, as both hoped and expected. For both normalizing alone and normalizing with caps at -2 and 2 there is more stability in being identified as focus than there is change, but for capping at -1.5 and 1.5 or -1 and 1, there is more change than stability.

The TAC was also shown the impact on gap measures of each of the four options, as show in figure 7. As can be seen from Figure 7, the distribution of composite achievement gap metrics remains relatively symmetrical when normalizing without caps, becomes slightly skewed to the right when normalizing and capping at -2 and 2, becomes increasingly skewed when capping at -1.5 and 1.5, and becomes extremely skewed when capping at -1 and 1.



Normalized without Caps

Normalized and Capped at -2, 2

Normalized and Capped at -1.5, 1.5

Normalized and Capped at -1, 1

Figure 7. Impact of normalizing and capping on the distribution of composite achievement gap.

The TAC was also shown the scatterplots in Figures 8-12 demonstrating the relationship between TTB percentile rank and composite gap measures for the original metric and the four modification options.

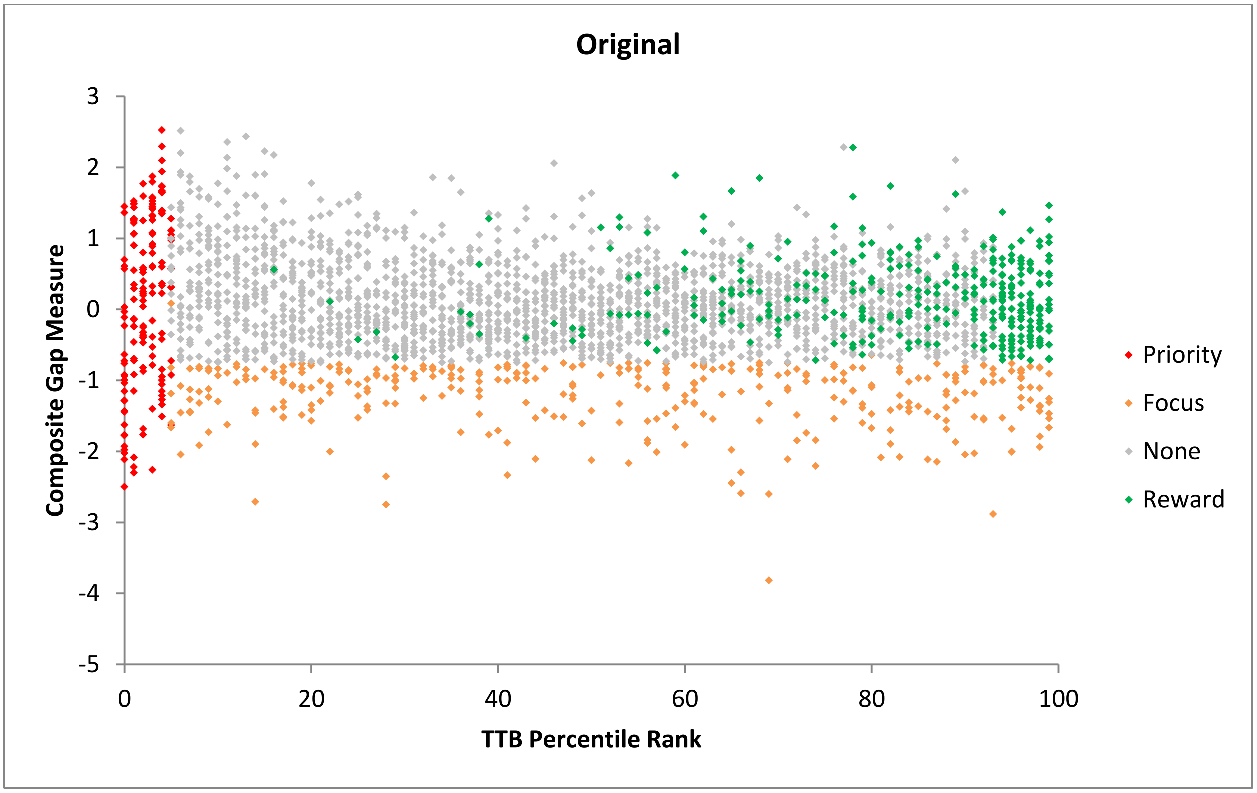


Figure 8. Original relationship between TTB percentile rank and composite gap.

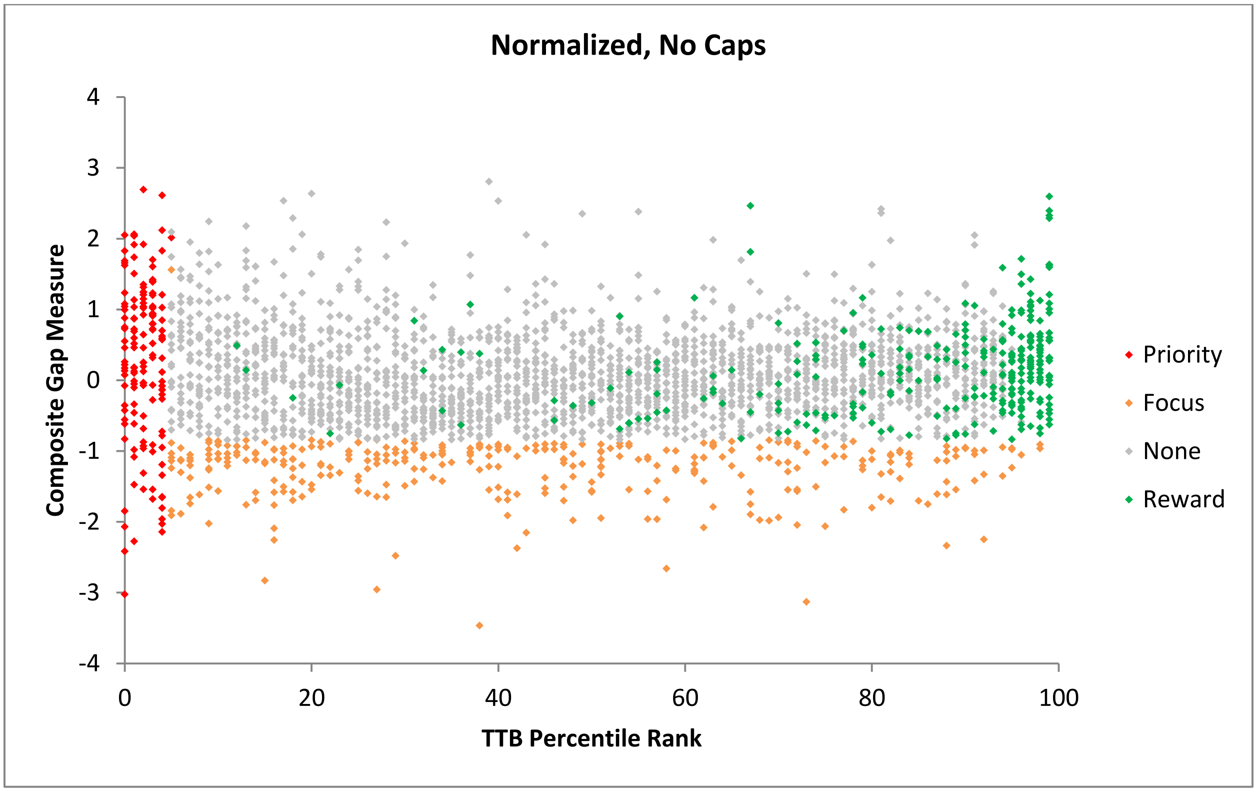


Figure 9. Relationship between TTB percentile rank and composite gap when normalizing alone.

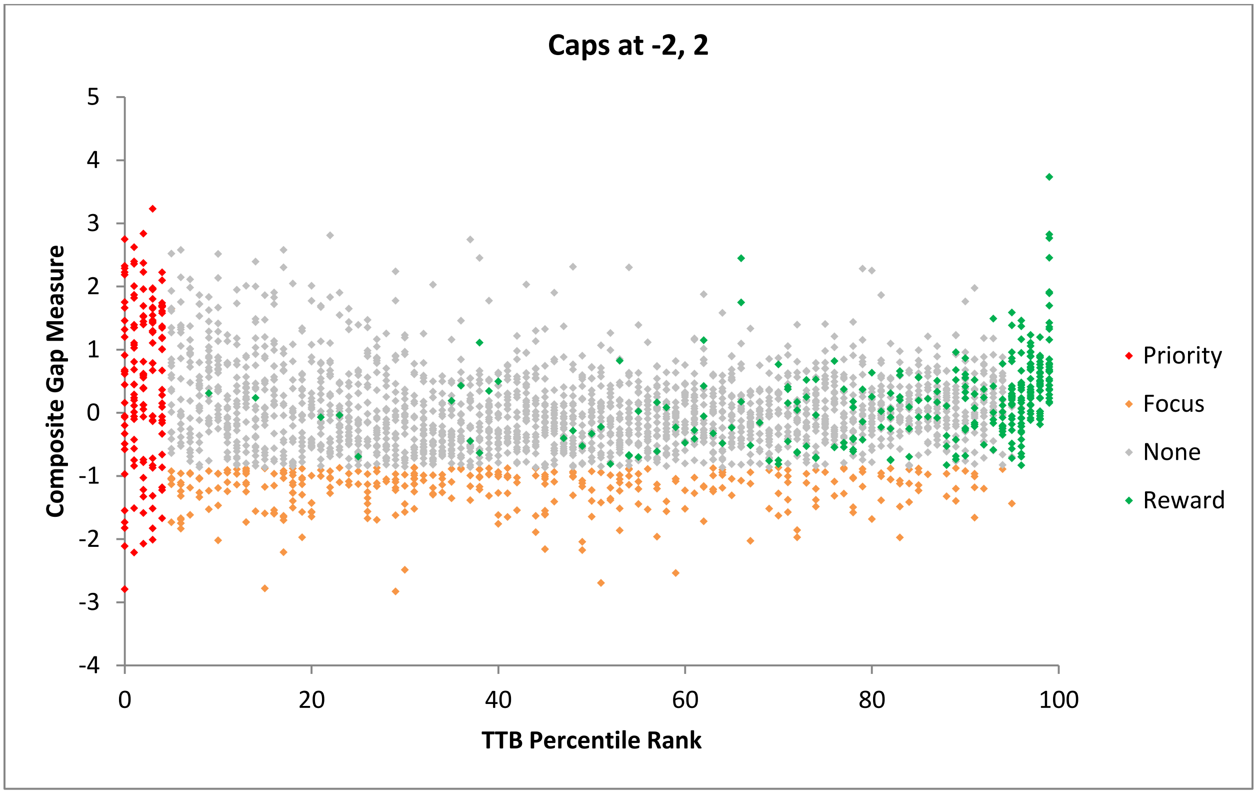


Figure 10. Relationship between TTB percentile rank and composite gap when normalizing and capping at -2, 2.

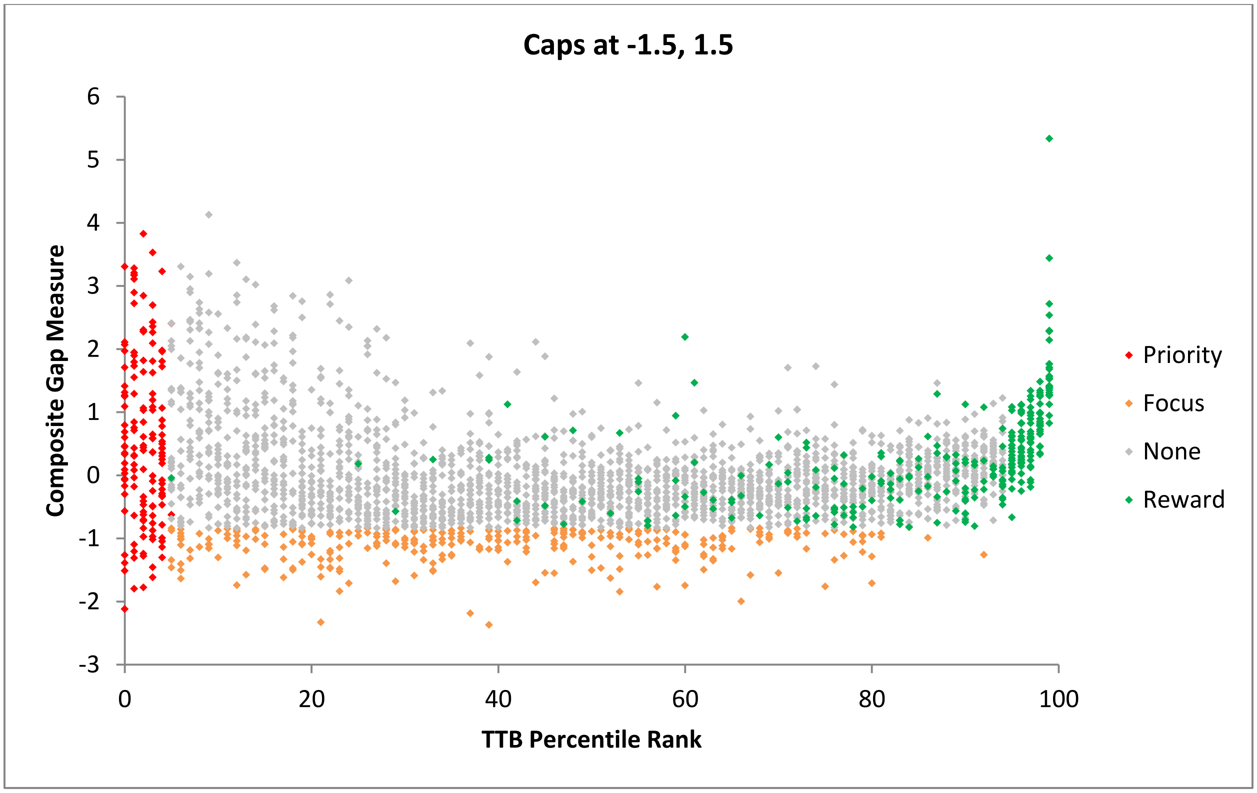


Figure 11. Relationship between TTB percentile rank and composite gap when normalizing and capping at -1.5, 1.5.

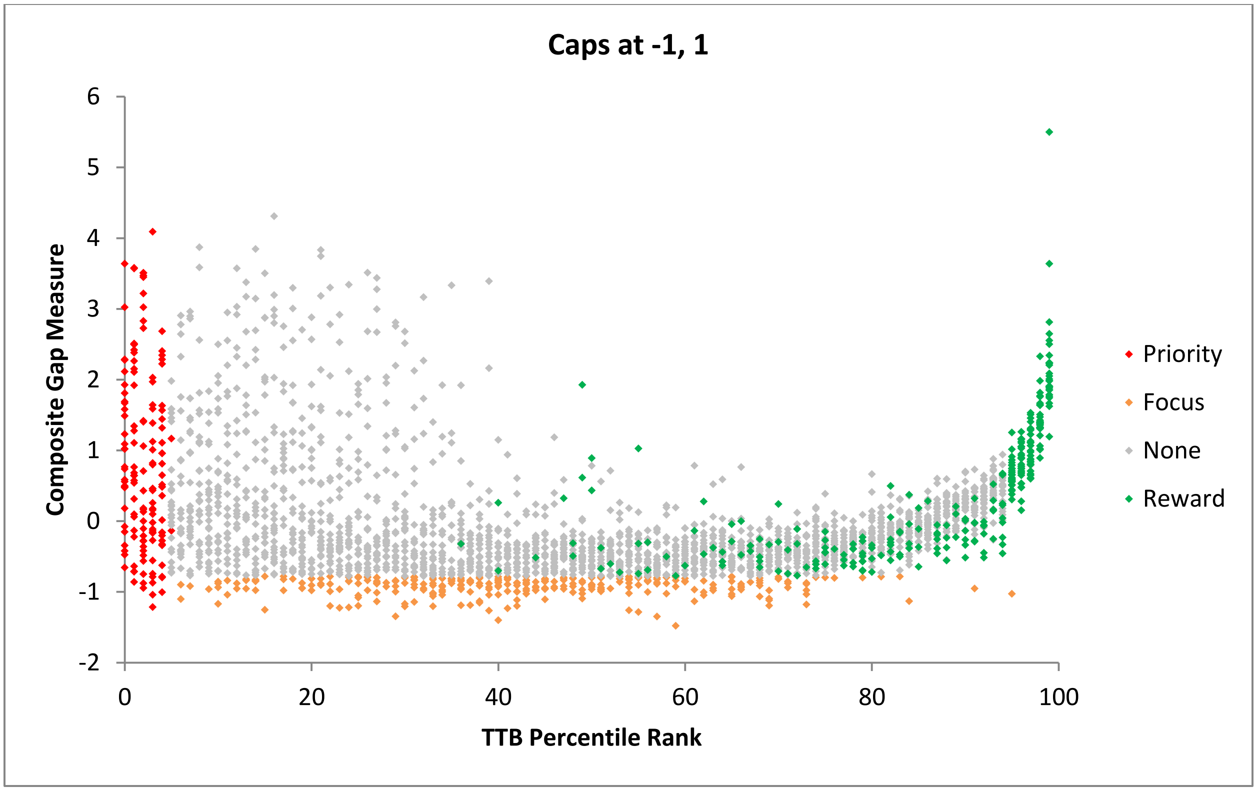


Figure 12. Relationship between TTB percentile rank and composite gap when normalizing and capping at -1, 1.

Figures 8-12 show the relationship between TTB percentile rank and composite gap, identifying priority, focus, and reward schools in each scenario. The impact of the choice of modifications is clear. Normalizing alone reduces the number of extremely high ranked schools that are identified as focus schools. Normalizing and capping at -2 and 2 increases that impact slightly, with no schools ranked above 95 identified as focus schools. Capping at -1.5 and 1 increases that impact markedly, with few schools above the 80th percentile identified as focus schools. Finally, capping at -1 and 1 identifies very few schools above the 75th percentile as focus schools.

The TAC was also shown the impact of the various choices on the relationship between percentage of students disadvantaged in a school and being identified as a focus school. Figures 13-17 show those relationships. Figures 13-17 show the relationships as well as identifying the priority, focus, or reward designation for each school. As can be seen from Figure 13, focus schools tended originally to be distributed throughout the range of economic disadvantage, with very poor schools often instead being identified as priority. Figure 14 shows that normalizing without caps results in fewer very well to do schools being identified as focus schools. Normalizing and capping at -2 and 2 slightly increases that impact. However, normalizing and capping at -1.5 and 1.5 significantly increases that impact. Finally, normalizing and capping at -1 and 1 results in focus schools being identified solely from schools in the middle range of economic disadvantage. This indicates that choosing to normalize and cap at -1 and 1 would result in identifying schools solely from those with the greatest economic diversity.

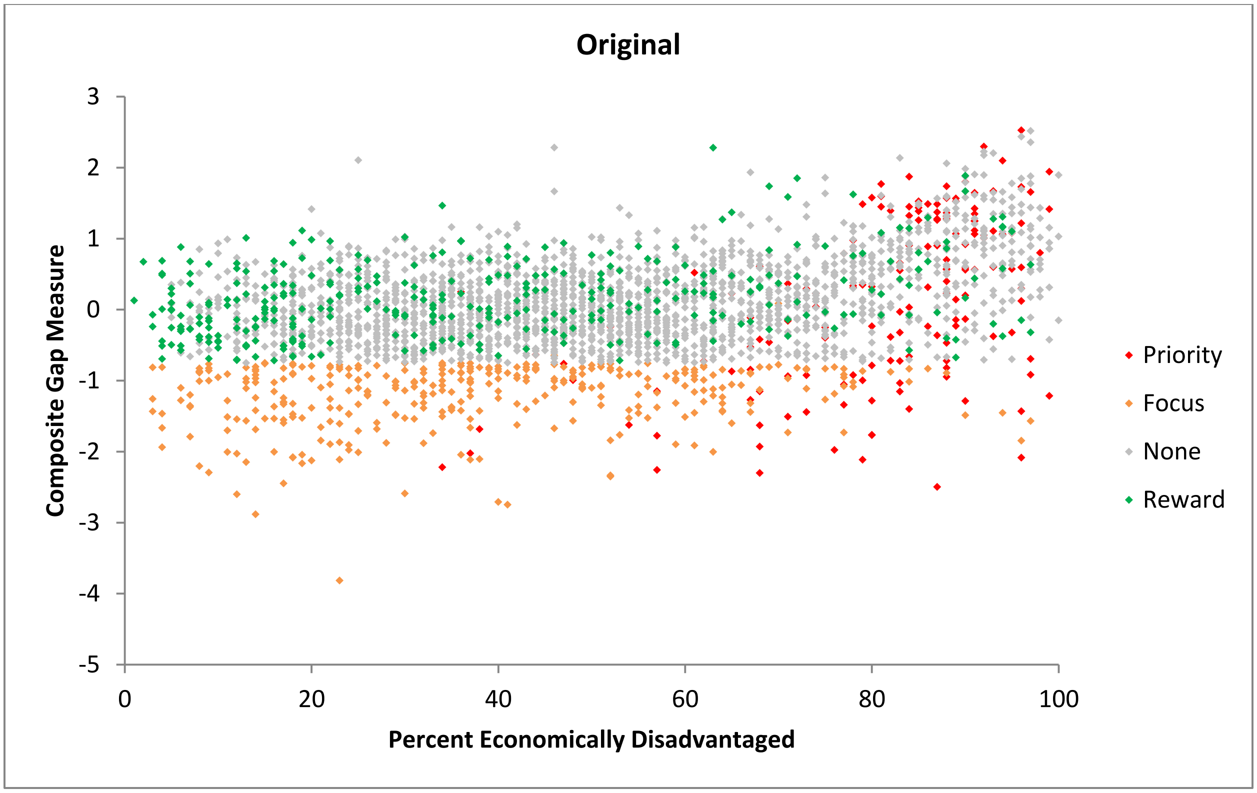


Figure 13. Original relationship between economic disadvantage and composite gap.

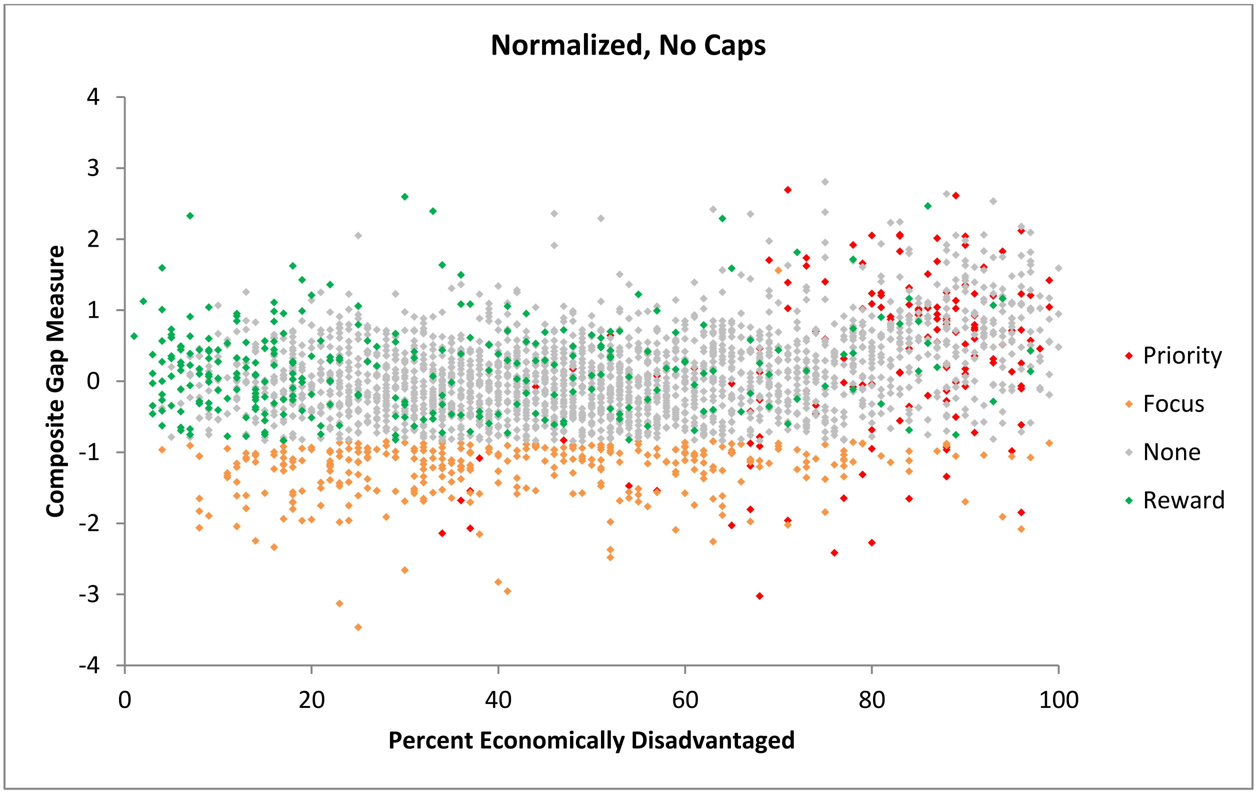


Figure 14. Relationship between economic disadvantage and composite gap when normalizing alone.

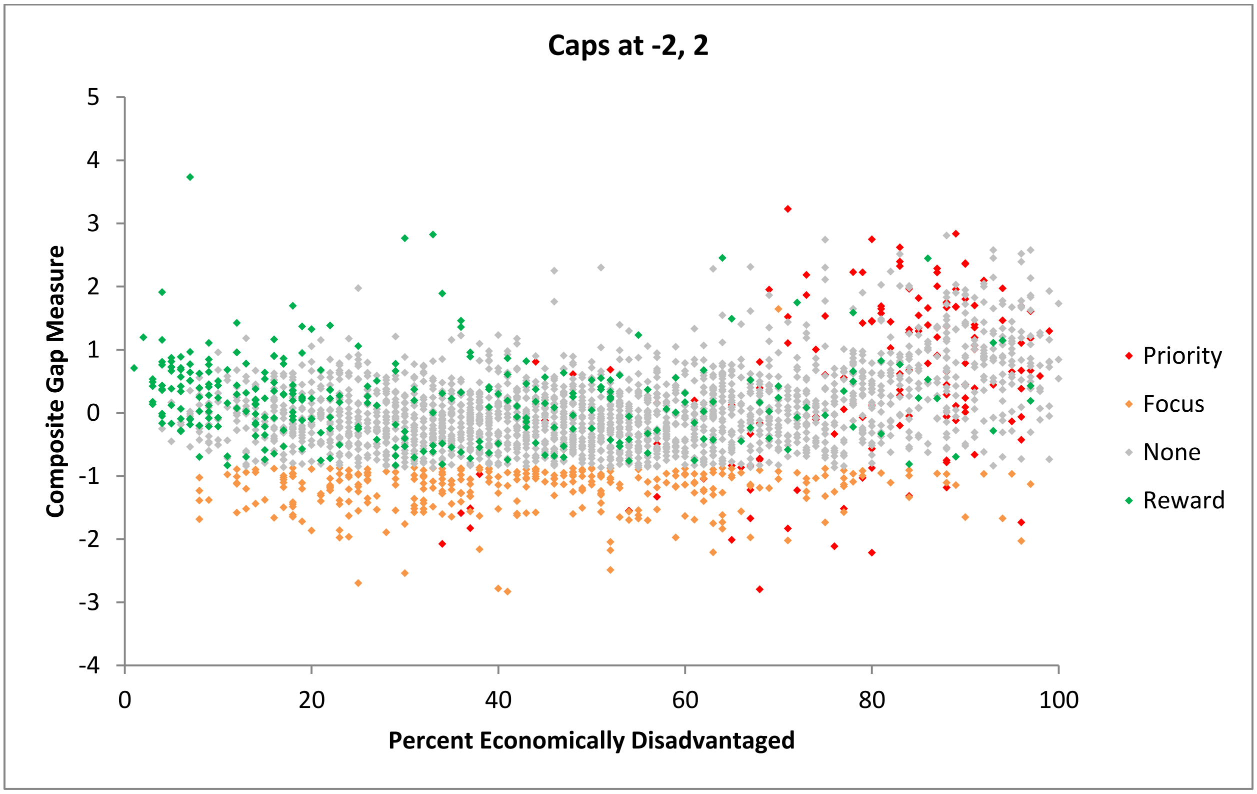


Figure 15. Relationship between economic disadvantage and composite gap when normalizing and capping at -2 and 2.

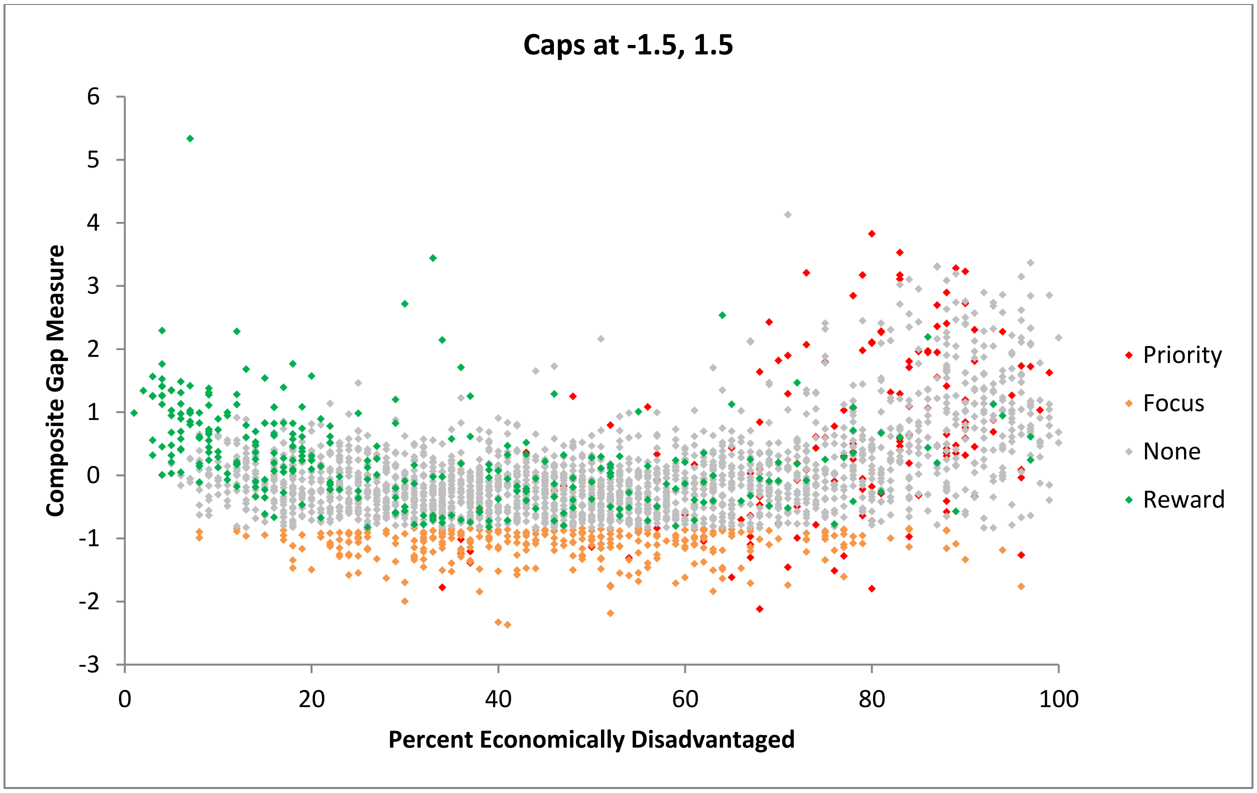


Figure 16. Relationship between economic disadvantage and composite gap when normalizing and capping at -1.5 and 1.5.

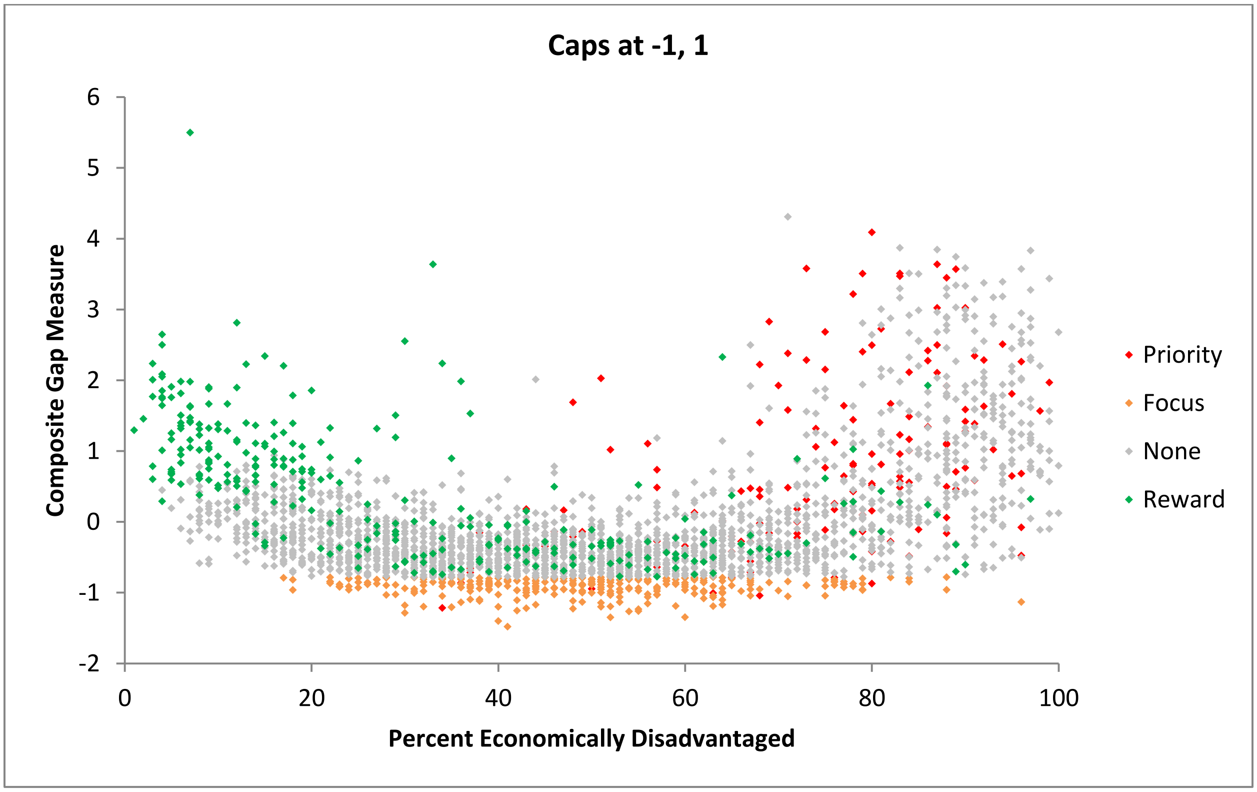


Figure 17. Relationship between economic disadvantage and composite gap when normalizing and capping at -1 and 1.

The TAC was also shown the impact of the various choices on the relationship between percentage of minority students in a school and being identified as a focus school. Figures 18-22 show those relationships, identifying the priority, focus, or reward designation for each school.

From Figures 18-22, it is clear that none of the options for modification has a large impact on the distribution of focus schools across the range of minority rates in schools.

Finally, the TAC was shown the relationship between composite achievement levels and composite gaps, for each of the five runs, as shown in Figures 23-27.

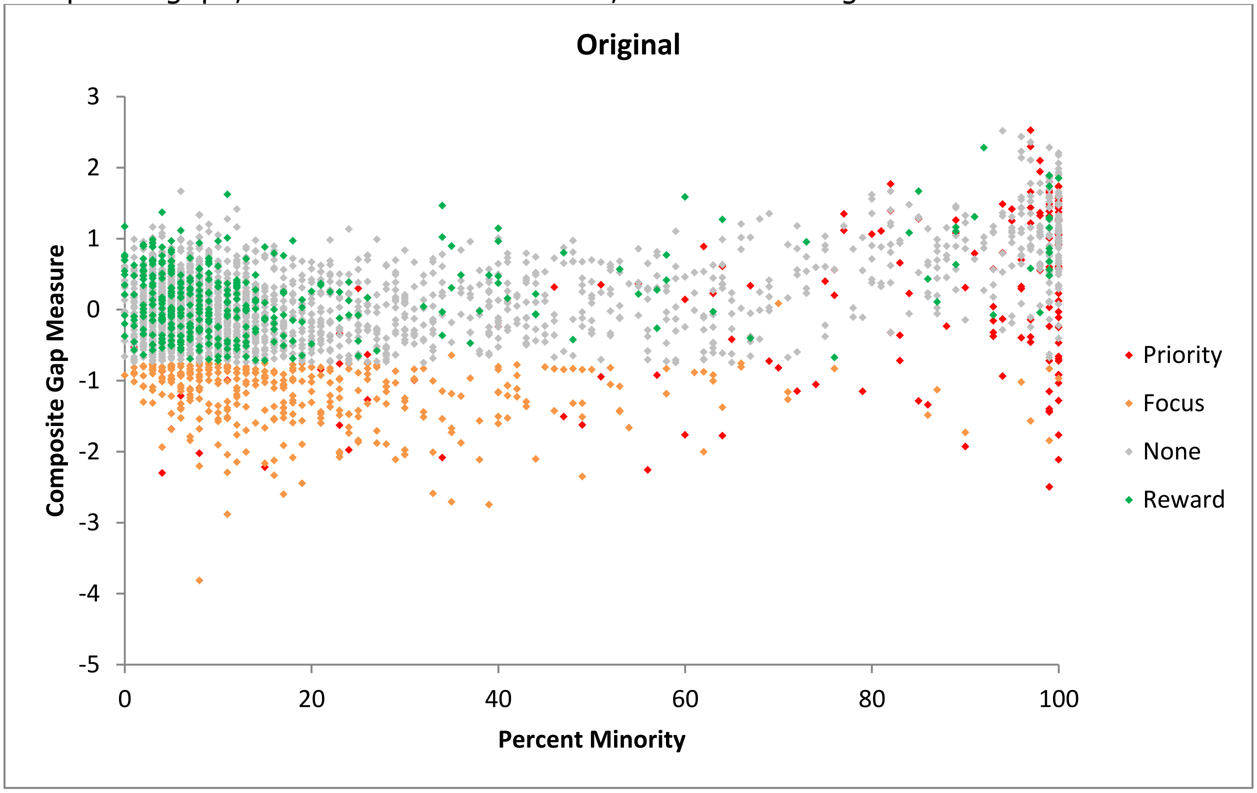


Figure 18. Original relationship between minority rate and composite gap.

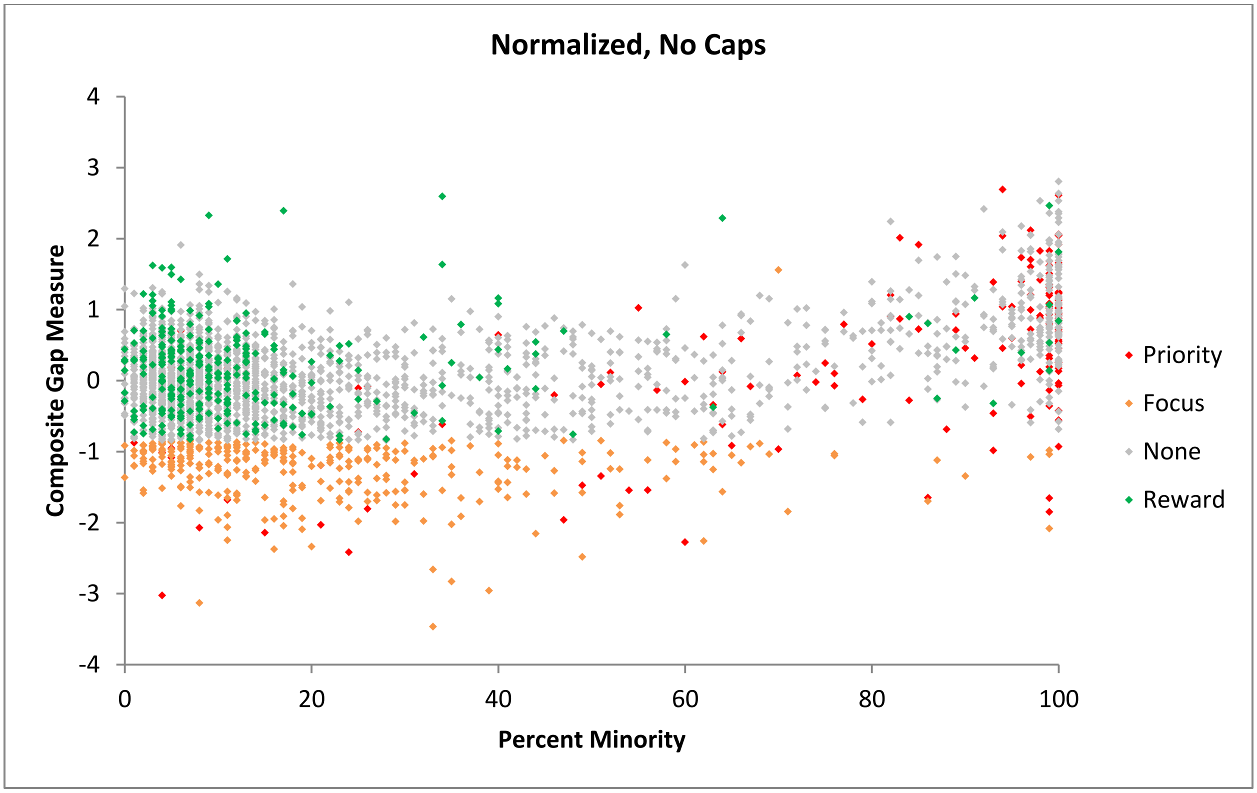


Figure 19. Relationship between economic disadvantage and composite gap when normalizing alone.

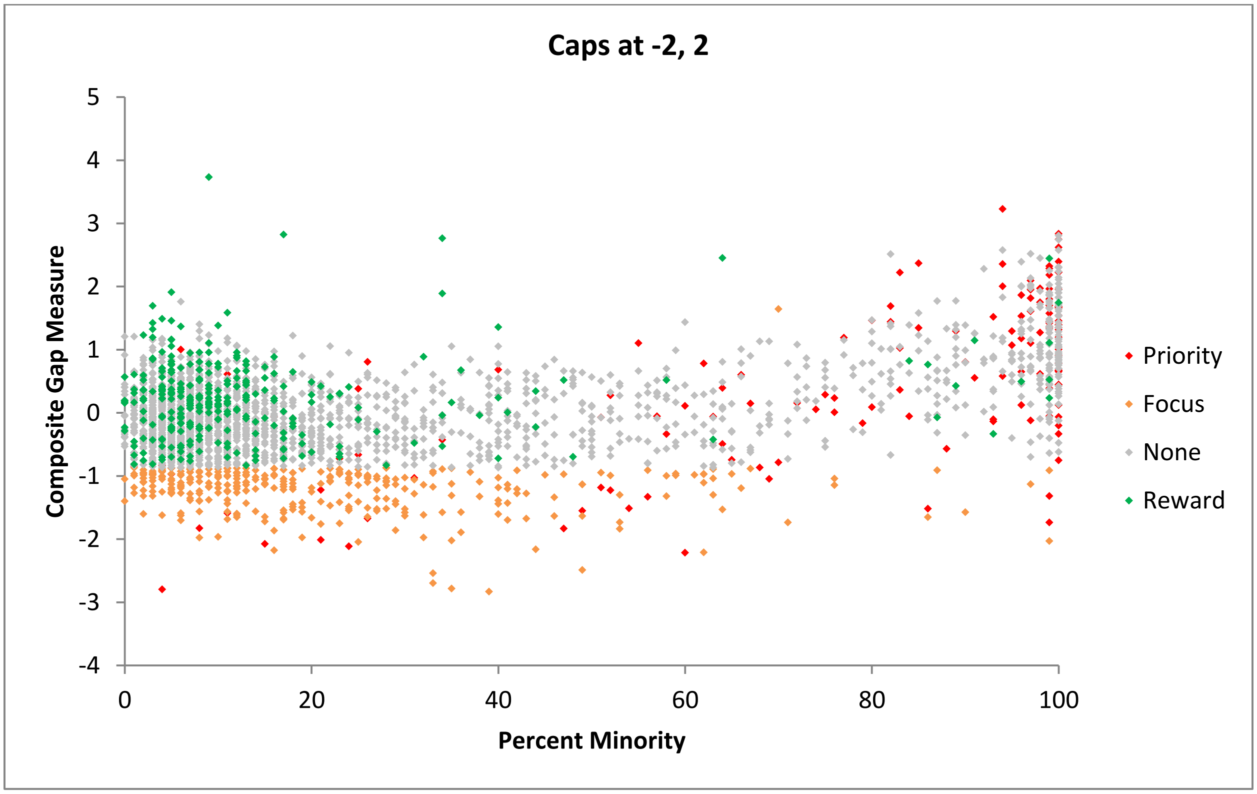


Figure 20. Relationship between economic disadvantage and composite gap when normalizing and capping at -2 and 2.

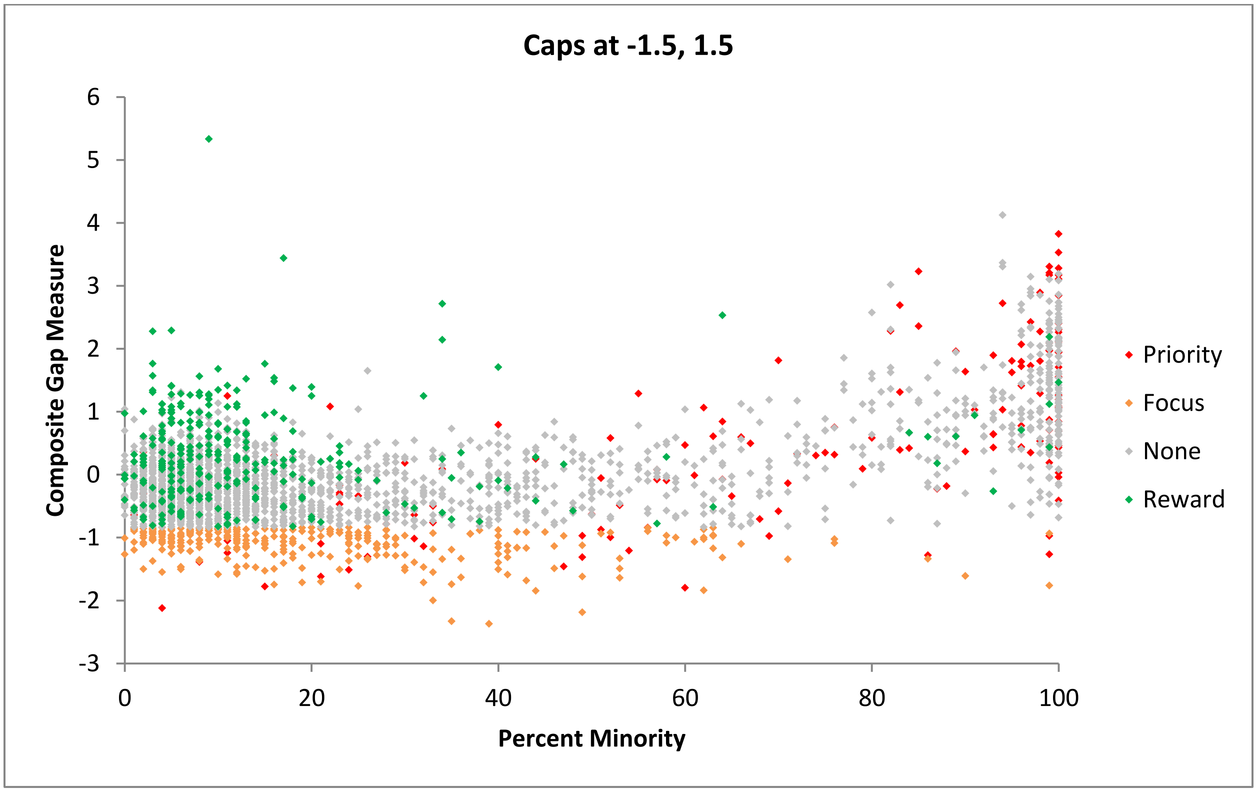


Figure 21. Relationship between economic disadvantage and composite gap when normalizing and capping at -1.5 and 1.5.

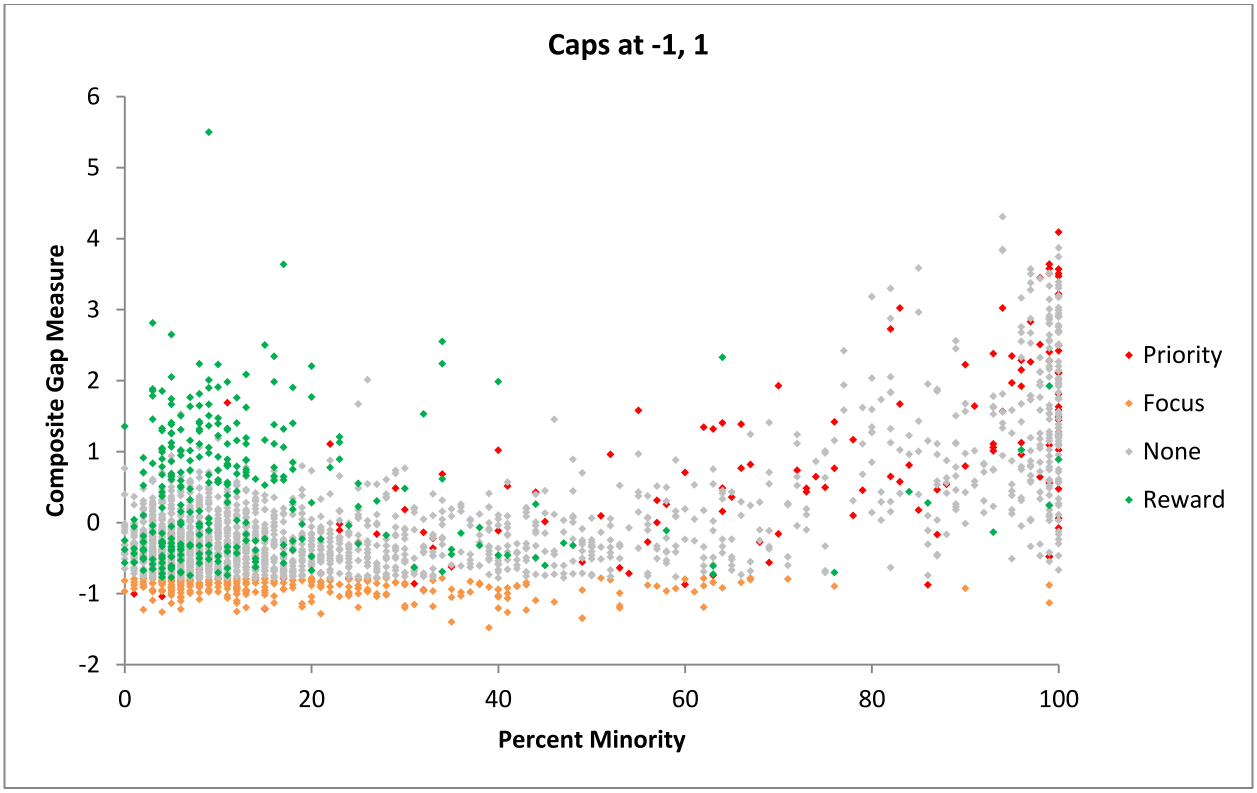


Figure 22. Relationship between economic disadvantage and composite gap when normalizing and capping at -1 and 1.

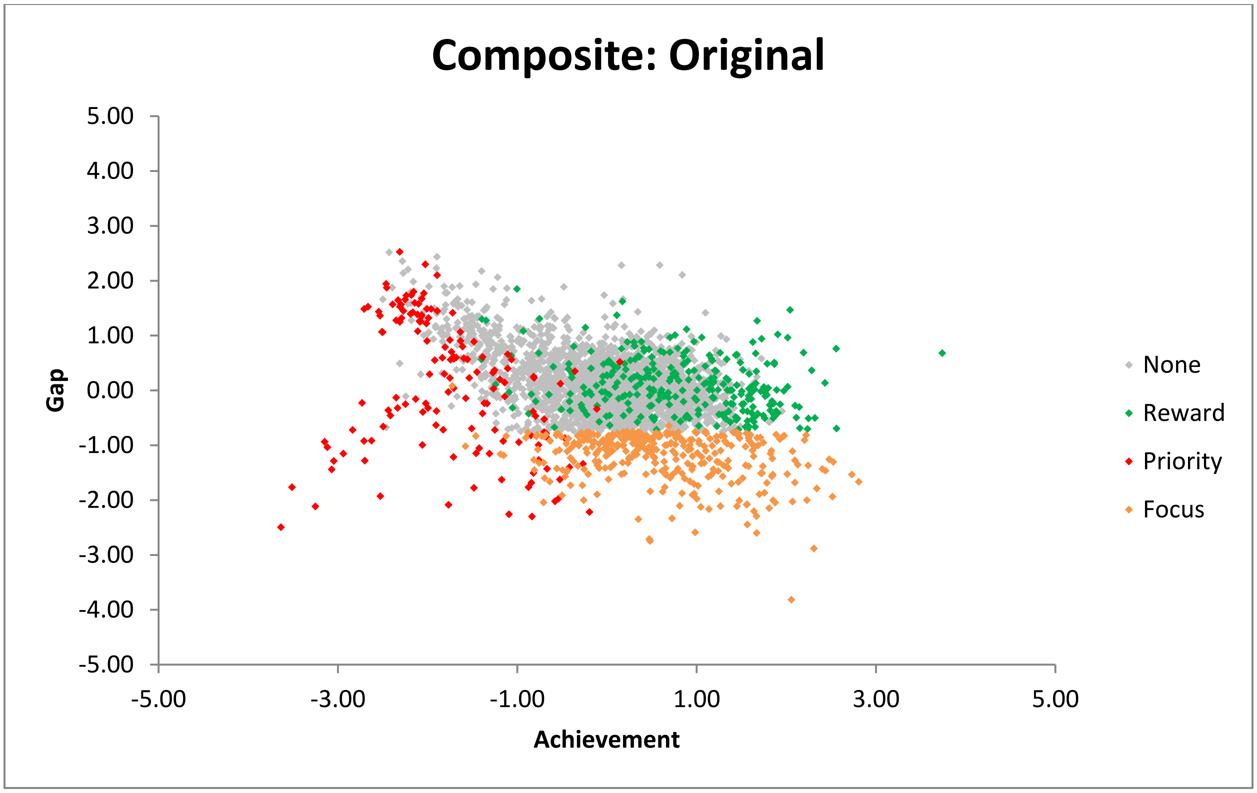


Figure 23. Original relationship between composite achievement and composite gap.

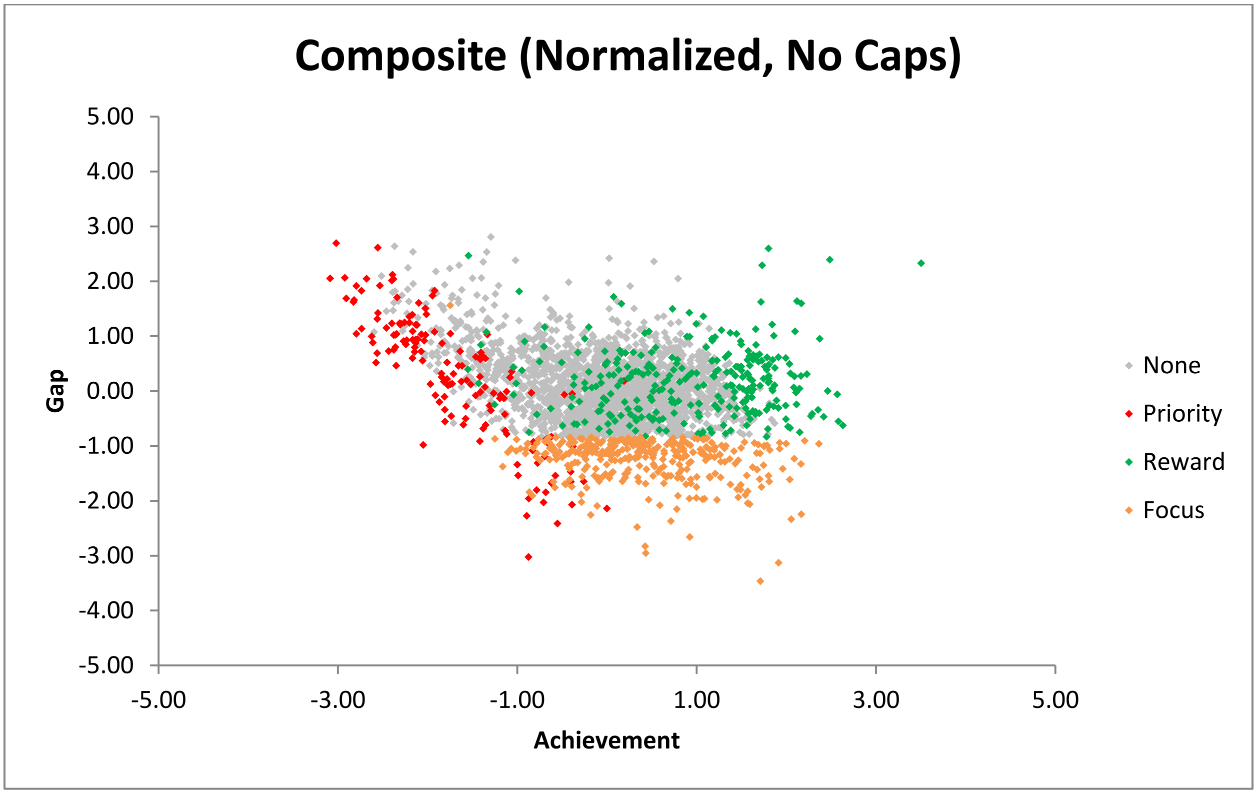


Figure 24. Relationship between composite achievement and composite gap when normalizing alone.

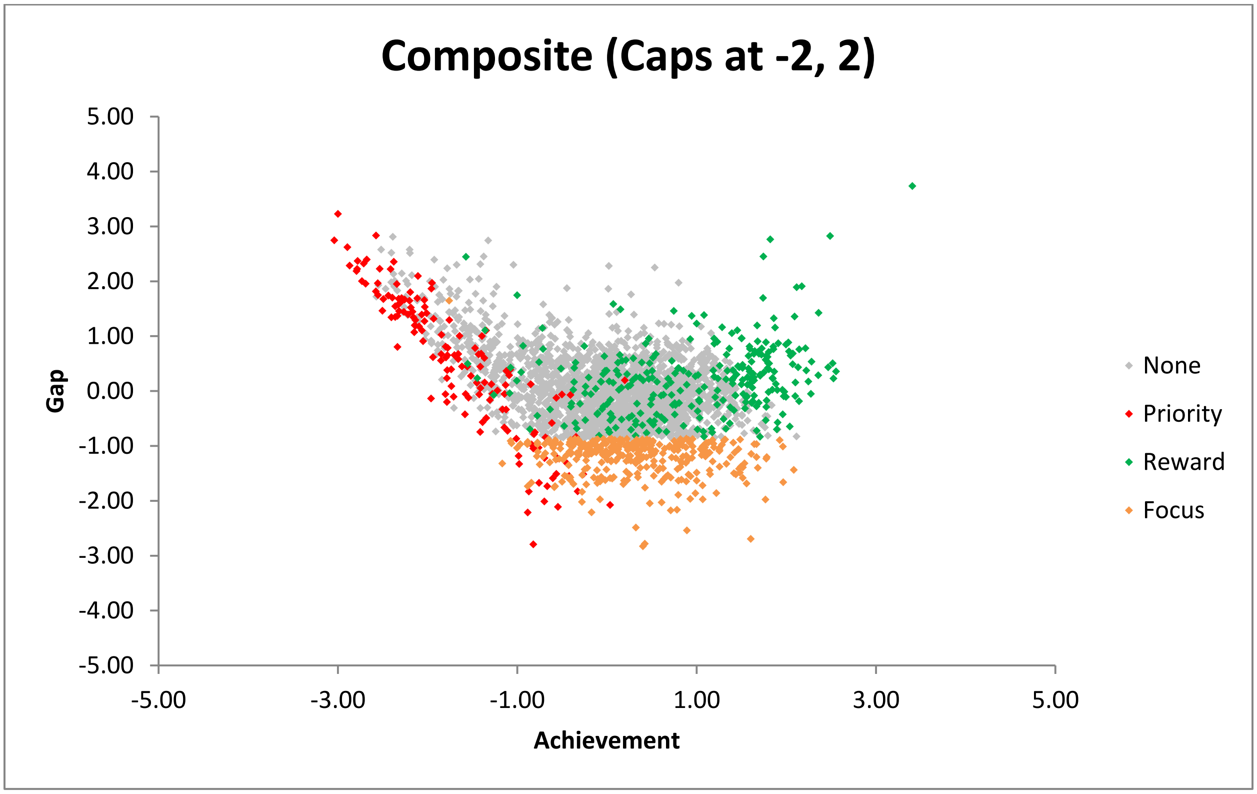


Figure 25. Relationship between composite achievement and composite gap when normalizing and capping at -2 and 2.

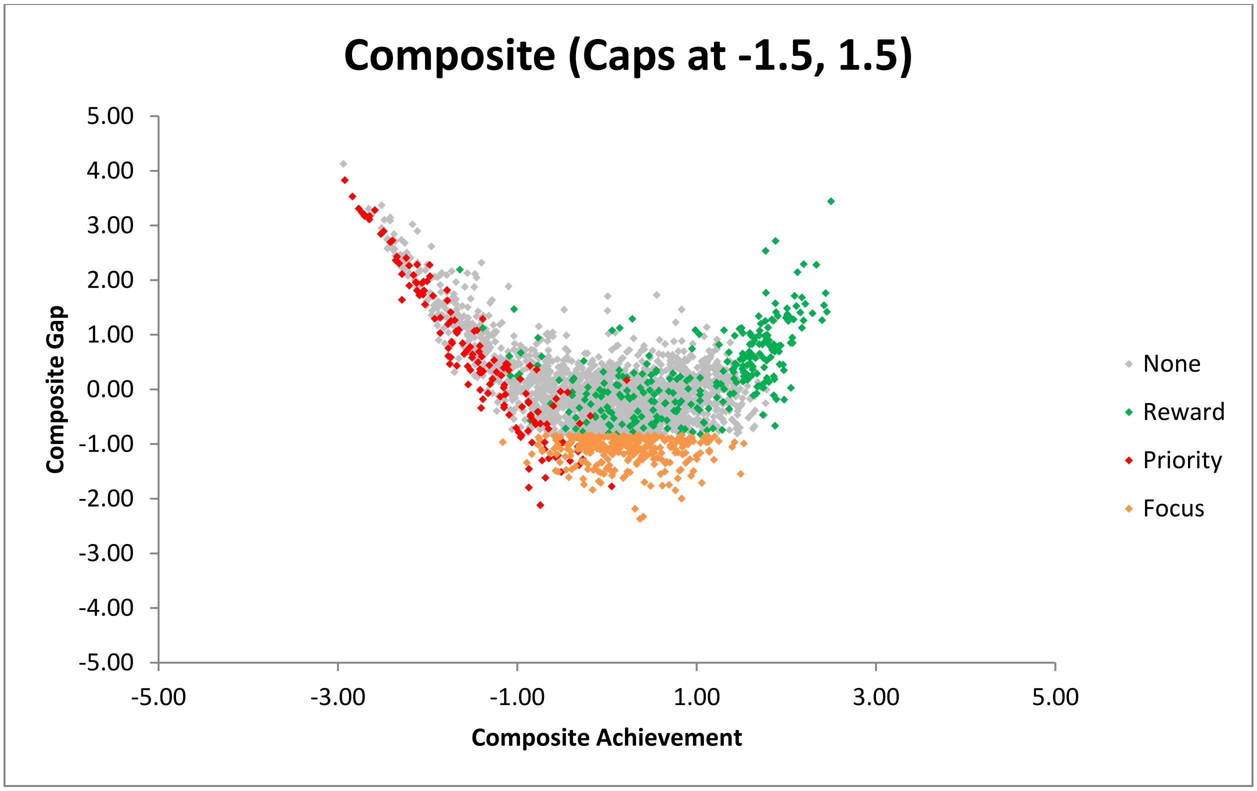


Figure 26. Relationship between composite achievement and composite gap when normalizing and capping at -1.5 and 1.5.

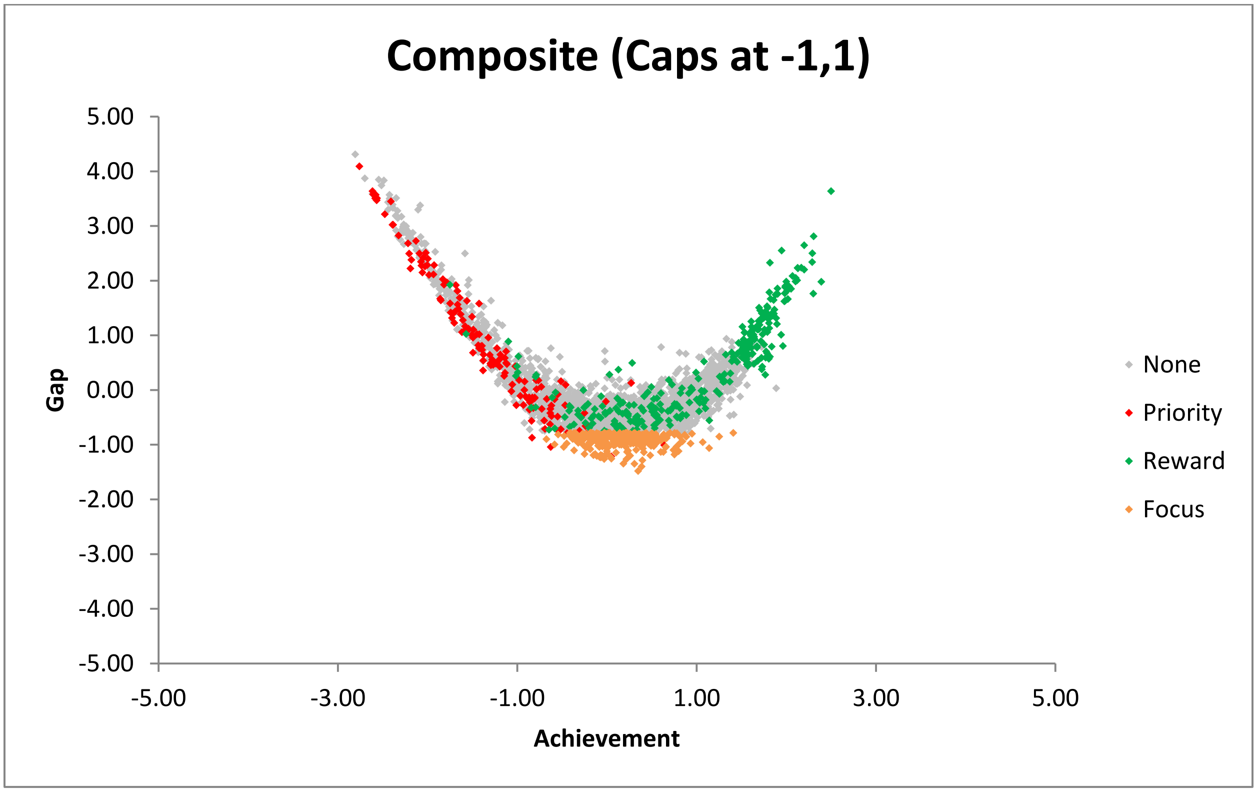


Figure 27. Relationship between composite achievement and composite gap when normalizing and capping at -1 and 1.

As can be seen in figure 23, the relationship between composite achievement and composite gap is negative for the lowest achieving schools, and relative unrelated for the remainder of schools. Normalizing along (figure 24) does not have a strong impact on the relationship, nor does normalizing and capping at -2 and 2 (figure 25). However, capping at -1.5 and 1.5 results in only schools in the middle range of achievement being identified as focus schools. Capping at -1 and 1 exaggerates that effect in that only schools from a small middle range of achievement are identified as focus schools.

The TAC recommended to BAA staff that in order to accomplish the object (to blunt the impact of outliers on focus identification), that the top to bottom metric should be modified by both normalizing student z-scores and by capping at least at -2 and 2. The TAC did indicate that capping at -1 and 1 would have a deleterious impact in terms of making the focus designation a proxy for middle levels of achievement and economic diversity. The TAC indicated that from a technical point of view the lower cap should lie somewhere between -2 and -1.5 and the upper cap should lie somewhere between 1.5 and 2, but the exact location of the caps is more a policy decision, and would be better deliberated upon by the BAA AC. The TAC also indicated that putting the information in some of the scatterplots into tables instead may help the BAA AC in interpreting the data.

BAA AC Meeting and Recommendations

The BAA AC was convened after the meeting with the BAA TAC. They were provided with the same information as the BAA TAC, plus the information in tables 4-8. Table 4 shows the average TTB rank of focus and non-focus schools and the maximum rank of a focus school under the five different methods of calculating the TTB metrics. As can be seen from Table 4, the average TTB rank of focus schools drops considerably when normalizing, with capping having a small effect. In addition, the average TTB ranking of non-focus schools increases slightly with normalizing and capping. Finally, the maximum ranking of focus schools decreases with normalizing and capping, indicating that fewer very highly ranked schools are identified as focus schools when normalizing and capping.

Table 4. Descriptive Statistics on TTB Rank.

Modification	Average TTB Rank of Focus Schools	Average TTB Rank of Non-Focus Schools	Max Rank of Focus Schools
Original	55	49	99
Normalized, no caps	41	51	98
Normalized, caps at -2, 2	39	51	95
Normalized, caps at -1,5, 1.5	39	52	92
Normalized, caps at -1, 1	42	52	95

Table 5 shows the number of priority schools by range of economic disadvantage, and table 6 shows the same for focus schools. It is clear from table 5 that normalizing has a minimal effect on the relationship between economic disadvantage and priority designation, with a slightly larger effect when adding in caps at -2 and 2. However, the impact of capping at -1.5 and 1.5 or -1 and 1 is considerable in that many more schools in the 26-50% range and the 51-75% range are identified as priority schools.

Table 5. Number of Priority Schools by Range of Economic Disadvantage

	Range of Economic Disadvantage			
Modification	<25%	25-50%	51-75%	>75%
Original	0	8	30	108
Normalized, no caps	0	8	32	105
Normalized, caps at -2, 2	0	9	35	101
Normalized, caps at -1,5, 1.5	0	12	46	88
Normalized, caps at -1, 1	0	15	50	81

Table 6 shows that normalizing reduces the number of schools identified as focus schools, and that capping reduces that number even further. The BAA AC found this to be a significant advantage. However, capping at -1.5 and 1.5 or at -1 and 1 does move many more focus schools into the middle ranges of economic disadvantage. Given that this results in identifying focus schools only from those that are the most economically diverse, the BAA AC found this to be a significant disadvantage.

Table 6. Number of Focus Schools by Range of Economic Disadvantage

	Range of Economic Disadvantage			
Modification	<25%	25-50%	51-75%	>75%
Original	118	134	87	19
Normalized, no caps	89	127	98	27
Normalized, caps at -2, 2	73	137	96	25
Normalized, caps at -1,5, 1.5	43	137	114	22
Normalized, caps at -1, 1	17	147	116	19

After discussion of the information presented and the issues surrounding the different options for modification, the BAA AC concurred with the BAA TAC recommendations of normalizing and capping at least to some degree. However the BAA AC indicated that capping at -2 and 2 was the preferable option in that it had minimal impact on the relationships between economic disadvantage and focus identification and between school achievement levels and focus identification. BAA AC did express concern that if caps other than -2 and 2 were implemented, priority identification would be limited to economically diverse schools and to schools in a small middle range of achievement.

However, the BAA AC members felt that while normalizing and capping at -2 and 2 would address the vast majority of problematic identifications of focus schools, there might still be a small number of schools whose bottom 30 groups are high performing enough to warrant their not being identified as focus schools. They recommended that BAA staff identify a reasonable threshold for the performance of bottom 30 groups that would exempt schools from being identified as focus schools if the bottom 30 group scored above that threshold. They also recommended that this threshold replace the good getting great exemption already in MDE's approved flexibility waiver.

BAA Identification of Bottom 30 Threshold to Exempt Schools from Being Identified as Focus Schools

BAA staff identified three possible thresholds for the bottom 30 subgroup for exempting schools from focus identification. These were:

1. Exempt schools from focus identification if their bottom 30 subgroup meets its scorecard target in at least two subjects and their TTB percentile rank is at least 75.
2. Exempt schools from focus identification if their bottom 30 subgroup scores higher than the overall state average in at least two subjects and their TTB percentile rank is at least 75.
3. Exempt schools from focus identification if their bottom 30 composite achievement is at or above the 90th percentile of composite achievement for bottom 30 subgroups.

While each threshold would exempt a similar small number of schools whose bottom 30 group is relatively high performing, each has different strengths. The strength of option 1 is that it is tied to the school scorecard. The strength of option 2 is that it is directly related to the criticisms many have leveled concerning the focus metric—that focus schools whose bottom 30 groups exceed the state average should not be considered focus schools. The strength of option 3 is that it is cleaner to implement. In evaluating the strengths of each option, it was clear that tying the threshold directly to one of the major criticisms of the metric was the most desirable.

Summary of Recommendations

Based on consultations with stakeholders, it is recommended that the top to bottom metric be modified in the following ways:

1. Normalizing student z-score distributions.
2. Capping student z-score distributions at -2 on the lower end and at 2 on the upper end.
3. Exempting from focus designation any school whose bottom 30 group scores at or above the state average in at least two subject areas.